Universiteit
Leiden
The Netherlands

Quantum methods for machine learning and classical dynamics

Barthe, A.M.

Citation
Barthe, A. M. (2026, March 20). *Quantum methods for machine learning and classical dynamics*. Retrieved from https://hdl.handle.net/1887/4297482

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/4297482

Note: To cite this publication please use the final published version (if applicable).

CHAPTER 4

Parameterized quantum circuits as universal generative models for continuous multivariate distributions

4.1. Introduction

Parameterized quantum circuits are the centerpiece of numerous approaches to machine learning on quantum computers, motivated by several near-term hardware limitations [4, 78]. The use of these models for supervised learning has been widely studied, for example in solving regression problems [121] where the goal is to assign continuous labels to data points. Variational algorithms have also been explored in so-called generative modelling tasks, where the objective is to generate new samples following a distribution that generated the training data [122].

A prominent example is the Quantum Boltzmann Machine (QBM) as introduced in [123], modelling distributions with discrete support, in which data is represented as the thermal state of an Ising model. In [124] the training based on the relative entropy of QBM for more general models is investigated, and it is shown that the training cannot be simulated with classical computers unless BPP=BQP. The QBM has been compared

The contents of this chapter have been published in [36].

in-extenso in [125] with another quantum generative model for discrete variables, the Quantum Circuit Born Machines (QCBM). In QCBM [45] a probability distribution over n-bit strings is stored in a n-qubit pure state. They may be trained to minimize the maximum mean discrepancy but also as generators in Generative Adversarial Networks (GANs) settings [93, 95, 126, 127], yielding the so-called Qu-GAN.

Going beyond distributions with discrete support, an approach to model distributions where the random variable can in principle take any value within a continuous interval has been introduced in [128]. In such models, the quantum circuit takes classical randomness as input and outputs expectation values, consequently, we call this model *expectation value samplers*. This model has been used as a quantum generator in the context of quantum generative adversarial networks (GAN), with applications including image generation [129, 130], high energy physics [131] and chemistry [132]. In all these applications the training is performed in the GAN setting.

While for QCBMs, the expressivity and universality have been clarified [44, 45], in contrast, expectation value sampling models are not so well understood. In particular, an interesting feature of expectation value sampling is that the dimension of the output is not inherently tied to the number of qubits used. In this chapter, we focus on the expressivity of the generators based on expectation value sampling depending on the number of qubits and the observable norm. In light of the numerous works [129–137] that successfully used expectation value samplers, this chapter aims to solidify the theoretical ground that supports them, by formally analyzing their expressivity. The formal analysis of the trainability of such models, while being a matter of critical importance, is not in the scope of this chapter.

The focus of this chapter is to show that parameterized quantum circuits are universal generative models and precisely characterize their expressivity. Our first main result is the existence of two universal families of expectation value samplers. We probe tight bounds relating to resource limitations, that is, necessary conditions on resources for expectation value sampling models to be universal. Specifically, we show that reaching universality for very high-dimensional distribution requires either a very large number of qubits or a very large number of measurements. This may serve as a backbone for future fine-grained resource cost analyses. As additional tools to analyze expressivity, we discuss choices of random variables, circuit encoding and observables, which we hope will guide future designs. Finally, we motivate the use of expectation value samplers with respect to other existing quantum generative models by providing a natural sampling task for these models.

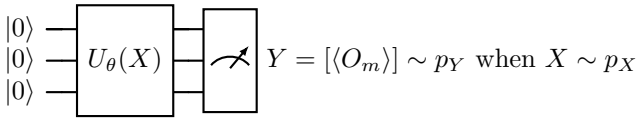

Figure 4.1.: Expectation value sampling model: a random vector is classically sampled. It is used to generate a random quantum state using a parameterized quantum circuit. The expectation values of fixed observables are returned as the output random vector.

4.2. Results

In this section, we introduce expectation value sampling models and define precisely the universality property for generative models. We state formally our first main result, namely the existence of two universal families of expectation value samplers. We then present our second main result, as the minimum resource requirements for expectation value samplers to be universal.

4.2.1. Expectation value sampling

The expectation value sampling procedure goes as follows. A random variable is classically sampled and used as input to a parameterized quantum circuit which specifies a random quantum state. The expectation values of fixed observables are measured and returned as another random variable. We illustrate this procedure in Figure 4.1, and provide a formal definition below.

Definition 4.1 (Expectation value sampling model)
An expectation value sampling model on n qubits is defined by $(U_\theta, \mathbf{O}, p_X)$, where $U_\theta : \mathcal{X} \subseteq \mathbb{R}^L \to \mathcal{U}(2^n)$ is a parameterized quantum circuit taking data as input and returning a matrix in the 2^n-dimensional unitary group $\mathcal{U}(2^n)$, $\mathbf{O} = (O_m)_{1 \leq m \leq M}$ is a vector of M observables, and $p_X : \mathcal{X} \subseteq \mathbb{R}^L \to \mathbb{R}$ is the input distribution. We define the associated mapping f as follows:

$$x \in \mathcal{X} \xrightarrow{f} \left(\langle 0 | \, U_\theta(x)^\dagger O_m U_\theta(x) \, | 0 \rangle \right)_{1 \leq m \leq M}. \tag{4.1}$$

The output of the model is a sample drawn from the distribution p_Y with $Y = f(X) \sim p_Y$ when $X \sim p_X$.

It is important to note that, in contrast to the quantum circuit Born

machine, the randomness does not come from the measurement process, but from the classical randomness provided as an input to the quantum circuit. Another difference is that expectation value samplers have continuous support (absolutely continuous random variables, see Supplementary Information). The central question of this chapter is whether such a model can generate any multivariate distribution, more precisely whether expectation value sampling is a universal generative model, according to the definition we will give in the next subsection.

4.2.2. Universal generative model family

In this subsection, we precisely define universal generative model families. We choose the Wasserstein distance to quantify the closeness of two distributions. For reasons we detail in the Supplementary Information, While the Wasserstein distance is impractical for training purposes, it is still a relevant metric in the analysis of the expressivity. A main reason is because it naturally arises as the mathematical concepts we use in this chapter are related to optimal transport. Other common losses are the Kullback-Leibler Divergence, which is not well suited to this chapter as it is not symmetric and not a true metric. Another frequently used loss is the Maximum Mean Discrepancy which is easier to compute, but relative to a chosen kernel, and therefore not an absolute property of the closeness between two distribution.

Definition 4.2 (Universal generative model family)
A generative model is a family of parameterized sampling procedures which enable the sampling from a corresponding set of M-dimensional probability density functions $\mathcal{P}(\mathcal{X})$ on $\mathcal{X} \subseteq \mathbb{R}^M$.

A generative model is called universal if for every probability density function q on $\mathcal{X}$ there exists a sequence $\{p_k | p_k \in \mathcal{P}(\mathcal{X})\}_{1 \leq k \leq \infty}$ such that it converges to q in the Wasserstein distance W.

$$W(q, p_k) \xrightarrow[k \to \infty]{} 0 \tag{4.2}$$

Importantly, universality is defined on a given support noted as $\mathcal{X}$ in Definition 4.2. For this chapter, we choose the support as the hypercube $\mathcal{X} = [-1, 1]^M$, because the first step of most machine learning pipelines is to rescale the data to fit on a given interval, usually $[-1, 1]$. Notably, choosing universality on a cubic support $[-1, 1]^M$ allows the expression of fully independent variables. Restricting $\mathcal{X}$ to smaller subsets would yield constraints on the dependence relationships expressible. For example,

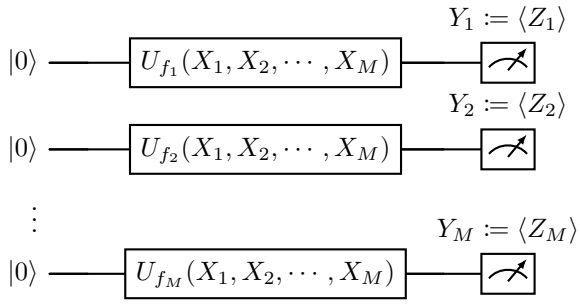

Figure 4.2.: *Product Encoding* Circuit as a universal generator yielding the random variable $Y = g(X)$ when $X \sim U([0,1]^M)$, by stacking circuits from [25] U_f approximating f, and defining $f_m \coloneqq \sqrt{(g_m + 1)/2}$.

proving the universality of a family of models for Dirichlet distributions would correspond to choosing $\mathcal{X}$ in the hyperplane where all variables sum up to one and are positive.

4.2.3. Two families of Expectation Value Samplers as Universal Generators

First, building on work by [25], we show that n-qubit expectation value samplers with constant observable norm are universal for M-dimensional distributions with constant support radius, for $M = n$.

Theorem 4.1
For any M, for all M-dimensional probability density functions p_Z with support included in $[-1,1]^M$, and for all accuracy $\varepsilon > 0$ there exists a M-qubit circuit U and set of M observables $\mathbf{O}$ with unit spectral norm $\|O_m\| = 1$ such that the expectation value sampling model $(U, \mathbf{O}, p_X)$ where p_X is the uniform distribution on $[0,1]^M$ yields a probability density function p_Y that is ε-close to p_Z in the Wasserstein distance.

We derive an explicit construction, illustrated in Figure 4.2, for this circuit, which yields product states, hence we name it "*product encoding*". It uses the same number of qubits as the dimension of the output distribution. This construction defeats one advantage of expectation value samplers, that is the dimension of the output not being directly linked to the number of qubits used. This raises the question of the existence of a more qubit-frugal family of universal circuits, in which the output dimension is (much) larger

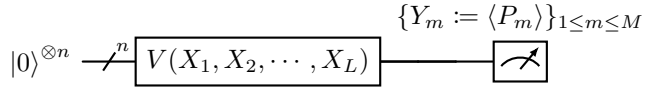

Figure 4.3.: *Observable Dense Encoding* Circuit as a universal generator, based on a universal state preparation circuit V, with each parameterized gate replaced by a circuit from [25]. $n = \log(M+1)$ and $P_m = 2M \left|m\right\rangle \left\langle m\right| - I$.

than the qubit number. We show the existence of such a family, at the cost of allowing for observables to have large norms. We formalize this in the following theorem and illustrate the explicit construction of the circuit in figure Figure 4.3 and call it "*observable-dense encoding*" expectation value sampler.

Theorem 4.2
For any M, for all M-dimensional probability density functions p_Z with support included in $[-1,1]^M$, and for all accuracy $\varepsilon > 0$ there exists a $n = \Theta(\log M)$-qubit circuit U taking L input variables and set of M observables $\mathbf{O}$ with spectral norm $\|O_m\| \in \Theta(M)$ such that the expectation value sampling model $(U, \mathbf{O}, p_X)$ where p_X is the uniform distribution on $[0,1]^L$ yields a probability density p_Y that is ε-close to p_Z in the Wasserstein distance.

In this subsection, we have provided *sufficient* conditions for expectation value samplers to be universal. In particular, we have shown the existence of two extremal families of parameterized quantum circuits that are universal generators on $[-1,1]^M$. There is a *product encoding* design, illustrated in Figure 4.2, with $n = M$ qubits and with unit norm observables (local Pauli), and an *observable-dense encoding* design, illustrated in Figure 4.3 with $n = \log(M)$ qubits and M norm observables (amplified probabilities of bitstrings). While one has a large number of qubits for a constant observable norm, the other has a logarithmic number of qubits but large observable norms. This hints that there might be a trade-off between the number of measurements and the number of qubits necessary to achieve universality. In the next section, we prove this is the case, by proving some *necessary* universality conditions.

4.2.4. Necessary conditions for universality

In this subsection, we prove two necessary conditions on resources for expectation value samplers to be universal. As previously mentioned, an appealing feature of expectation value sampling models is that the dimension of the output vector is *a priori* independent of the number of qubits n, unlike in the case of quantum Born machines where each qubit corresponds to exactly one binary random variable. In particular, we can imagine using just a single qubit with an arbitrary number of observables O_m to generate an M-dimensional random vector. However, it is obvious that in this case, the random variables corresponding to each observable cannot all be fully independent. Indeed, any observable can be expressed as a linear combination of the three Pauli matrices. Therefore the distribution output by a single qubit expectation value sampler will have at most three degrees of independence, and for $M > 3$ it is impossible to reach universality because it is impossible to approximate e.g. the 4-dimensional uniform distribution. Extending this reasoning to several qubits, we find the first necessary condition, that the dimension of the target dimension has to be lower or equal to the dimension of the space of observables. We formalize this in the following lemma.

Lemma 4.1
For an n-qubit expectation value sampling model $(U_\theta(x), \mathbf{O}, p_X)$ to be able to approximate any distribution with support in $[-1,1]^M$ to any accuracy $\varepsilon > 0$, it is necessary that $M \leq 4^n - 1$.

The second necessary condition for an expectation value sampling model to be universal on distributions with support in $[-1,1]^M$ can be derived using a combination of Holevo's bound found in [138] and Chernoff bound. We formalize it in the following theorem.

Theorem 4.3
For an n-qubit expectation value sampling model $(U_\theta, \mathbf{O}, p_X)$ to be able to approximate any distribution with support in $[-1,1]^M$ to any accuracy $\varepsilon > 0$ with respect to the Wasserstein distance, it is necessary that for every $m \leq M$, both equations are satisfied:

$$\lambda_{\min}(O_m) \leq -1 + \varepsilon \text{ and } \lambda_{\max}(O_m) \geq +1 - \varepsilon\,, \tag{4.3}$$

$$n \in \Omega\left(\frac{M(1-\varepsilon)^2}{\|O_m\|^2}\right)\,. \tag{4.4}$$

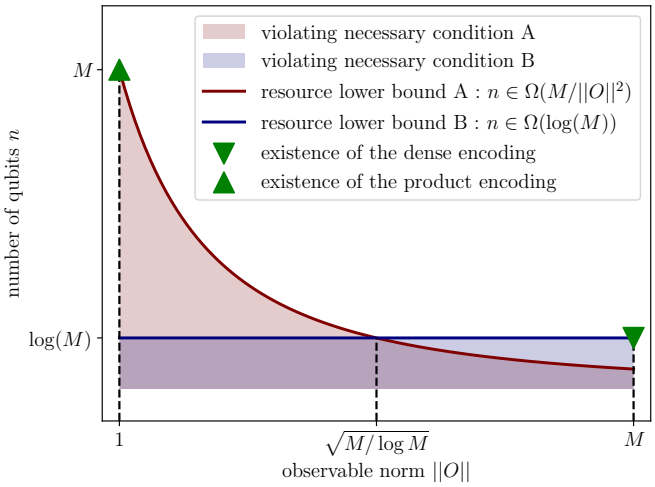

Figure 4.4.: Visual summary of results. We show the asymptotic use of resources for expectation value samplers to reach universality for a M-dimensional target distribution: the necessary conditions, as well as the existence of families as in Figure 4.2 and in Figure 4.3.

with $\lambda_{\min/\max}(O)$ *returning respectively the minimum and maximum eigenvalues of observable* O*.*

The combination of both previous necessary conditions is the second central result of this chapter. It formalizes that even if expectation value sampling models may output arbitrary large dimensional distributions, in practice their expressivity is limited by the number of qubits and observables. It is a natural question to ask whether the universal families we found previously saturate the above necessary conditions. We illustrate such considerations in Figure 4.4. For target distributions with exponentially large dimensionality, the two universal families are (almost) asymptotically optimal.

We may conjecture that there exists a family of (Pareto) optimal circuits, balancing between observable norms and qubit numbers, with varying *encoding densities* in between the two extremal ones. In practice, the observable spectral norm relates to the number of measurements required to approximate it up to an additive accuracy. This highlights a trade-off between the number of qubits and the number of measurements, or space versus time complexity, which we formalize in the next subsection.

4.2.5. Trade-off: qubits vs measurements

To estimate an expectation value up to a desired constant additive error, the number of measurements required is proportional to the norm of the observables. Therefore, large observable norms require a large number of measurements.

Lemma 4.2 (informal)
To guarantee that the M-dimensional distribution coming from an expectation value sampler performing T measurements is ε-close in the Wasserstein distance to the ideal shot-noise-free distribution, it is sufficient that

$$T \in \Theta\left(\frac{M\|O\|}{\varepsilon^2}\right). \tag{4.5}$$

Therefore the necessary conditions we derived previously ultimately highlight a trade-off between the number of qubits and the number of measurements. We show that to be universal for very high-dimensional distributions on $\mathcal{X}$, expectation value samplers need either a very large number of qubits or a very large number of measurements. This is in contrast with the initial intuition that expectation value samplers may generate high-dimensional distributions independently from the number of qubits.

4.3. Discussions

The arguments we present to derive resource lower bounds necessary for achieving universality in expectation value samplers (EVS) are driven by the number of independent variables of the target distribution. Because our definition of universality allows for full independence relationships across all dimensions, the number of independent variables is equal to the dimension of the target distribution. A more refined perspective emerges when considering families of distributions with inherent dependencies and allows for more economical sampling strategies than the ones requiring full independence as presented in Figure 4.4. Two concrete examples serve to emphasize this point. First, consider the family of distributions that are M-dimensional but have their support confined within a 2-dimensional linear subspace. In this case, using the expectation value sampler described in [25] with the observables $\sqrt{2}X, \sqrt{2}Z$, we achieve a universal generator on a single qubit for this 2-linear family, independent of the

target distribution dimension M. This economy of representation underscores the capacity of EVS to exploit structural sparsity effectively. Next, consider M-dimensional Dirichlet distributions, characterized by non-negative coordinates summing to unity. Here, the observable dense encoding circuit, configured with $n = \log M$ qubits, enables universality with the raw probabilities as observables (Figure 4.3). This model achieves universal representation with exponentially fewer qubits compared to the general setting that allows for full independence, leveraging that the Born rule naturally gives rise to Dirichlet distributions. These findings suggest that EVS models are particularly well-suited for generating highly correlated distributions. The Dirichlet case is illustrative, as its normalization condition aligns seamlessly with that of quantum states, lending itself naturally to the expectation value sampling paradigm.

Building on this, we discuss the impact of design choices in EVS, specifically the selection of observables, the encoding $U(x)$, and the input distribution p_X, all of which are detailed further in the Supplementary Information. The choice of observables directly influences the subspace of the distribution spanned by the EVS. Furthermore, regarding the encoding structure $U(x)$, we realize that, analogously to the way parameterized quantum circuits may have an exact finite Fourier decomposition, expectation value samplers may have an exact finite Generalized Polynomial Chaos Expansion. In this context, the choice of the uniform distribution as an input distribution arises naturally, while in contrast to classical GANs, a Gaussian distribution yields undesirable inductive bias. Thanks to this proximity between the supervised and the generative context, results on the expressivity of quantum reuploading models [35] transfer naturally to the encoding circuits of expectation value samplers. In addition, as the Expectation Value samplers effectively learn a function in full analogy to how it is done in supervised learning, it is clear that much of the current discussion on the trainability of parameterized circuits for supervised learning, and of more general variational algorithms will apply to the Expectation Value samplers. It will have the same bottlenecks of overly expressive architectures [18], but also the various classes of new techniques that mitigate these by designs that balance trainability and dequantizability [21, 22, 139], and other more practical methods which mitigate trainability problems [59].

While we characterized important properties of expectation value samplers, we did not address the question of whether it is a good idea to use them. Expectation value samplers cannot be proven to be a path for certain types of quantum advantage as directly as is the case in Quantum Circuit Born Machines. QCBM are well suited for quantum use cases by

design as they rely on the Born rule for generation of samples, and so correspond to distributions obtained by a measurement of a genuinely quantum state. In contrast, expectation value samplers rely on classical randomness and, rather than requiring sampling from the full distribution of quantum measurements, are defined around expectation values only. Thus the hardness of simulating distributions from expectation value samplers does not connect straightforwardly to any hardness-of-sampling results established in the domain of quantum supremacy results [44]. Nonetheless, expectation value samplers still consist of genuinely quantum computations, and arguments for non-simulatability can be made. Assuming BQP is not in BPP, there exists no polynomial time algorithm that takes a classical description of an arbitrary expectation value sampler A as an input and outputs a sample from a distribution that is epsilon close to that of the output of A. Take as an example the case where a hard-to-simulate circuit does not depend on input data. This yields a Dirac delta distribution for which there exist classical samplers to efficiently sample from it, however, such classical samplers cannot be easily found based on the classical description of the expectation value sampler. It may be possible to construct stronger advantage arguments where we find the existence of an expectation value sampler such that its output distribution cannot be sampled by any polynomial-time randomized Turing machine (subject to standard assumptions). This raises the question: do expectation value samplers also correspond to any natural quantum task?

A natural domain of application for EVS emerges in many-body physics, particularly in the study of disordered systems. We propose the example of spin glasses, which with interaction strengths $J_{i,j}$ taken at random, are governed by the Hamiltonian:

$$H(J) = \sum_{\langle i,j \rangle} J_{i,j} S_i S_j - M \sum_i S_i \,. \tag{4.6}$$

In this model, the interaction strengths $J_{i,j}$ are taken at random. Consider the quench dynamics of such a system, initialized as the ground state of the Hamiltonian without any interaction $H(0)$, that is the zero state. The interactions are instantly turned on for a fixed time and a set of interactions is taken at random. Then one may compute the expectation value of a set of observables of interest such as local spin, or local correlators, measured using several copies of the same realization of interaction strengths. This data is inherently quantum, and the trotterization of this time evolution constitutes a solution in the form of an expectation value sampler where the output can be made arbitrary close to the target

distribution. We exemplify such a circuit for a chain geometry on four qubits in Figure 4.5. Using two registers with different evolution time, one could even extract two-time correlation function $C(t,t') = \sum_i \langle S_i(t) S_i(t') \rangle$, which in turn give precious information about the Edwards-Anderson parameter, an order parameter for spin glasses.

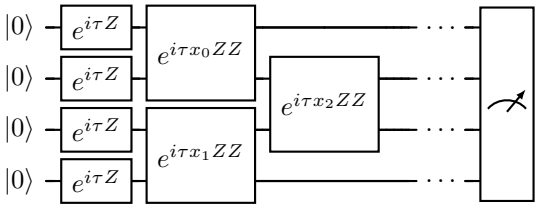

Figure 4.5.: Expectation value sampler as the trotterization of the spin glass Hamiltonian in Equation (4.6).

4.4. Methods

The proofs of all theorems are available in the Supplementary Information, but we provide high-level ideas in this section. First, we highlight a core concept in this chapter, variable transformation, and how we use it to prove universality. Subsequently, We explain the main steps of the proof and construction of the two universal families as a high-level summary of Section II of the SI .

4.4.1. Random Variable Transformation

A core concept in this chapter is that of random variable transformation. In this subsection, we introduce it and provide some of the associated fundamental properties.

The first step in the analysis of expectation value samplers is the observation that they are processes that map an input random vector (parameterizing the quantum circuit) to an output random vector (the expectation values of a set of observables). The literature on optimal transport and measure theory [140] states that for every pair of absolutely continuous random variables of the same dimension, there exists a mapping to transform one into the other.

Lemma 4.3
For every pair of absolutely continuous probability density functions $p_X \in \mathcal{P}(\mathcal{X} \subseteq \mathbb{R}^{\mathcal{M}})$ and $p_Y \in \mathcal{P}(\mathcal{Y} \subseteq \mathbb{R}^{\mathcal{M}})$, there exists a mapping $f : \mathcal{X} \to \mathcal{Y}$, that maps $Y \sim p_Y$ to $X \sim p_X$ as $Y = f(X)$.

We give intuition on how to construct this mapping to be bounded piece-wise continuous in the Supplementary information. Many generative modelling systems including Generative Adversarial Networks, Variational Auto-Encoders and normalizing flows rely on this idea to generate arbitrary distributions. In particular, the initial distribution is chosen to be simple, and then by altering the mapping applied to this random input, we obtain a rich spectrum of possible output distributions.

Then, sufficient conditions for a family of mappings to yield a universal generative model, in the sense of Definition 4.2, are well-known, and expressed in terms of the universality of mappings themselves. From [141], it is sufficient for a family of mappings to be dense in the set of all monotonically increasing functions in the pointwise convergence topology to yield a universal generative model. We explain the difference between pointwise topology and uniform topology in the Supplementary Information. Since these are sufficient conditions and monotonic functions are included in all functions, the following holds.

Lemma 4.4 ([141])
If a family of mappings $\mathcal{G} = \{g : \mathcal{X} \subseteq \mathbb{R}^M \to \mathcal{Y} \subseteq \mathbb{R}^M\}$ is dense in the set of all functions in the pointwise convergence topology, then this family of mappings $\mathcal{G}$ together with a probability density function p_X with non zero support on $\mathcal{X}$ yields a universal generative model family on $\mathcal{Y}$ (cf Definition 4.2).

In classical machine learning, such a notion for universality is common. It has been proven for several families of mappings in the context of normalizing flows, which are mappings with the additional property of being invertible: generic triangular mappings [140], neural networks mappings [142] and polynomial mappings [143].

In the next section, to prove the universality of expectation value samplers, we make explicit the connection between universal mapping families and the known results on the universality of parameterized quantum circuits as supervised learning models.

4.4.2. Universality proofs

In recent literature, several universality properties of parameterized quantum circuits have been proven. In [25], a family of single qubit quantum circuits with an increasing number of layers is proven to be universal in the uniform sense for continuous multidimensional input functions as complex coordinates of the quantum state in the computational basis.

We extend this existing result to fit our needs, as follows. First, we modify universality results from functions as coordinates in the computational basis to functions as the expectation value of an observable. Then we broaden the universality of quantum reuploading models to some discontinuous functions by relaxing the required strength of convergence. More precisely we go from the uniform density in bounded continuous functions to the pointwise density in bounded piece-wise continuous functions. Finally, by stacking M universal circuits, we extend universality to multivariate output functions. All these extensions of [25] together yield the theorem below.

Theorem 4.4
For every natural number M, for every mapping $f \in \mathcal{B}([0,1]^M \to [-1,1]^M)$, there exists a sequence of sets of M single qubit quantum circuits and unit norm observables (indexed by k).

$$\{(U_{k,m} : [0,1]^M \to \mathcal{U}(2), O_{k,m})_{1\leq m\leq M}\}_{1\leq k\leq\infty} \tag{4.7}$$

such that the sequence of functions $\{g_k\}_{1\leq k\leq\infty}$ defined as

$$g_{k,m}(x) = \langle 0| U_{k,m}(x)^\dagger O_{k,m} U_{k,m}(x) |0\rangle \tag{4.8}$$

converges pointwise to f, where $\mathcal{B}$ is the set of piecewise continuous functions and the norm of observable is the spectral norm.

The theorem above shows that there exists a family of M-qubit circuits with unit norm observables that yield a family of functions that is pointwise dense in the set of bounded piece-wise continuous functions. This matches the sufficient conditions of Lemma 4.4 for mappings to yield a universal generative model. This yields that there exists a family of expectation value sampling models as defined in Definition 4.1 and illustrated in Figure 4.2 that is universal in the sense of Definition 4.2.

For the *observable dense encoding* we follow a similar strategy, but instead, each output variable is encoded as the overlap between the output state and each computational basis state. With the normalisation of the quantum states the observables are projectors on computational basis

states, amplified by a factor proportional to the dimension of the target distribution. More details are provided in the Supplementary Information.

4.5. Appendix

4.5.1. Universality Definition

In this appendix, we spend a bit more time justifying our choice for the definition of universality and relating it to concepts of interest. The definition of universality for generative models hinges on the notion of closeness between two random variables. In the case of generative modelling, *convergence in distribution*, a common concept in probability theory, is often sought. For example, the central limit theorem precisely states the average of any L independent random variables with mean μ and variance Σ converges in distribution to the normal distribution $\mathcal{N}(\mu, \Sigma/L)$.

Definition 4.3 (Universal generative model family)
A generative model is a family of parameterized sampling procedures which enable the sampling from a corresponding set of M-dimensional probability density functions $\mathcal{P}(\mathcal{X})$ on $\mathcal{X} \subseteq \mathbb{R}^M$.

A generative model is called universal if for every probability density function q on $\mathcal{X}$ there exists a sequence $\{p_k | p_k \in \mathcal{P}(\mathcal{X})\}_{1 \leq k \leq \infty}$ such that the sequence of random variables $X_k \sim p_k$ converges in distribution to $X \sim q$. Equivalently, this means that the sequence of cumulative distribution functions of p_k, which we call P_k, converges pointwise to the cumulative distribution function of q, which we call Q.

$$\forall x \in \mathcal{X}, \lim_{k \to \infty} P_k(x) = Q(x). \tag{4.9}$$

In this chapter, we will mostly consider distributions with finite support, because this is the case in most practical real-world problems. In particular, their probability density functions are integrable functions and convergence in distribution implies convergence in the first Wasserstein distance W_1 [144]. For completeness, we recall below the definition of the Wasserstein distance, also known as the Earth Mover's Distance, that we use in the context of this chapter.

Definition 4.4 (Wasserstein distance)
The k-th Wasserstein distance between two probability density functions p

and q on $[-1,1]^M$ is defined as:

$$W_k(p,q) = \left(\inf_{\gamma \in \Pi(p,q)} \int_{\mathbb{R}^2} \|x-y\|^k d\gamma(x,y) \right)^{\frac{1}{k}}, \tag{4.10}$$

where $\Pi(p,q)$ is the set of couplings of p and q, and $\|\cdot\|$ denotes the Euclidean distance. The parameter $k \geq 1$ determines the so-called order of the Wasserstein distance.

For the rest of the chapter, we will only consider the first-order Wasserstein distance ($k = 1$) and therefore simply refer to it as the Wasserstein distance.

Importantly, universality is defined on a given support noted as $\mathcal{X}$ in Definition 4.3. For this chapter, we choose the support to be the hypercube $\mathcal{X} = [-1,1]^M$, because the first step of most machine learning pipelines is to rescale the data to fit on a given interval. Note that the length of the hypercube can be rescaled, by rescaling the norm of the observables. Finally, since any distribution with infinite support but finite moments can be arbitrarily approximated by a distribution with finite support, this means that if we allow the observable norm to scale, we can also approximate any distribution with finite moments (but perhaps infinite support). We make this explicit below.

We consider a distribution with probability density function p_Y with infinite support $\mathbb{R}^M$ and finite moments. The goal is to find a sequence of random variables with bounded support probability density function that converges pointwise to p_y. We define the sequence of random variables $\{Y_k\}_{1 \leq k \leq \infty}$ whose probability density functions p_{Y_k} are proportional to that of Y on the hypercube $[-k-k_0, +k+k_0]^M$, where k_0 is the first integer such that Y has non-null support on the hypercube. This sequence converges in distribution to Y.

4.5.2. Universality Proofs

In this appendix, we provide the proofs for theorems 1 and 2 that state the universality of two families of EVS. First, we provide more details on random variable mapping which is the core concept of the proof in Section 4.5.2. For both proofs, we use Theorem 4.5 from [25]. For the first theorem, in Section 4.5.2 we modify universality results from functions as coordinates in the computational basis to functions as the expectation value of an observable. In Section 4.5.2 we broaden the universality of quantum reuploading models to some discontinuous functions by relaxing

the required strength of convergence. More precisely we go from the uniform density in bounded continuous functions to the pointwise density in bounded piece-wise continuous functions. Finally, by stacking M universal circuits, we extend universality to multivariate output functions. All these extensions of [25] together yield Theorem 4. With this theorem, we show that the *product encoding* circuit satisfies the universal mapping property, and therefore, using the concept of random variable mapping, yields a universal generative model. For the *observable dense encoding* in Section 4.5.2 we follow a similar strategy, but instead, each output variable is encoded as the overlap between the output state and each computational basis state. With the normalisation of the quantum states the observables are projectors on computational basis states, amplified by a factor proportional to the dimension of the target distribution.

Constructive mapping between a random variable and the uniform distribution

Let us consider an absolutely continuous random variable Y with probability density function p_Y on a bounded set $[a, b]^M$. We recall a definition of an absolutely continuous variable below.

Definition 4.5
A random variable X is said to be absolutely continuous if its cumulative distribution function (CDF) can be expressed as the integral of a non-negative function, known as the probability density function (PDF).

This excludes for example Dirac deltas. In what follows we construct an invertible mapping to transform the uniform random variable X into this random variable $Y = [Y_k]$. We call $G_1 : [0, 1] \to [a, b]$ the cumulative distribution function of the marginal of Y_1. It is invertible and we define F_1 as its inverse,

$$Y_1 = F_1(X_1), X_1 = G_1(Y_1). \tag{4.11}$$

Next we consider the marginal of Y_2 conditioned by Y_1, we define the cumulative distribution $G_2 : [a, b]^2 \to [0, 1]$,

$$G_2(y_1, y_2) = P(Y_2 = y_2 | Y_1 = y_1). \tag{4.12}$$

It is invertible with respect to Y_2,

$$G_2^{-1}(Y_1, X_2) = Y_2 \iff G_2(Y_1, Y_2) = X_2. \tag{4.13}$$

We define $F_2 : [0,1]^2 \to [a,b]$ as follows,

$$F_2(X_1, X_2) = G_2^{-1}(F_1(X_1), X_2). \tag{4.14}$$

We have $Y_2 = F_2(X_1, X_2)$. Continuing this process iteratively for all coordinates, it is possible to fully define the invertible mapping $F = [F_k]$ with inverse $G = [G_k]$ such that $Y = F(X)$. This is the essence of triangular mapping in [140]. In addition, this mapping is bounded and piece-wise continuous, because it is composed of inverse cumulative distribution functions of absolutely continuous variables on bounded support.

From state coordinate universality to expectation of observable universality

We start by recalling Theorem 4 from [25], which proves that the following circuit is universal. It has L layers, $\theta \in \mathbb{R}^{(M+2)\times L}$ parameters and for $x \in \mathbb{R}^M$ is defined as

$$U_\theta(x) \coloneqq \prod_{l=1}^{L} R_y(\theta_{0,l}) \left(\prod_{m=1}^{M} R_z(x_m \theta_{m,l}) \right) R_z(\theta_{M+1,l}). \tag{4.15}$$

Theorem 4.5 (from [25])
For any pair of functions and real number

$$(f \in \mathcal{C}([0,1]^M \to [0,1]), \phi \in \mathcal{C}([0,1]^M \to [0,2\pi)), \varepsilon > 0)$$

There exists a one qubit circuit $U : [0,1]^M \to \mathcal{U}(2)$ *s.t.*

$$\forall x, \left| \langle 1| U(x) |0\rangle - f(x) e^{i\phi(x)} \right| < \varepsilon. \tag{4.16}$$

In this chapter, $\mathcal{C}$ is the set of continuous functions. This theorem yields the universality of functions embedded in a quantum state in the uniform sense. In the context of expectation value sampling, we are interested in the universality of function as the expectation value of a unit norm observable, captured by the following theorem.

Theorem 4.6
For any function $g \in \mathcal{C}([0,1]^M \to [-1,1])$ *and for any* $\varepsilon > 0$*, there exists a one qubit circuit* $U(x) : [0,1]^M \to \mathcal{U}(2)$ *and an observable* O *with unit spectral norm* $\|O\| = 1$ *s.t.*

$$\forall x, \left| \langle 0| U^\dagger(x) O U(x) |0\rangle - g(x) \right| < \varepsilon. \tag{4.17}$$

Proof: We are given an arbitrary function $g \in \mathcal{C}([0,1]^M \to [-1,1])$ and $\varepsilon > 0$. We define the function $f = \sqrt{\frac{g+1}{2}}$ which is well defined on $\mathcal{C}([0,1]^M \to [0,1])$. We apply Theorem 4.5 to $(f, \phi = 0, \varepsilon/4)$ and get a circuit U that yields a state close to

$$|x\rangle = \sqrt{1 - f(x)^2}\,|0\rangle + f(x)\,|1\rangle\,. \tag{4.18}$$

The Z expectation value of the above state is

$$\langle x|\, Z\, |x\rangle = 2f(x)^2 - 1 = g(x). \tag{4.19}$$

We are now going to prove that the expectation value is close to the target function g. First, we note that the square function is 2-Lipschitz on $[0,1]$ and therefore for every pair of real numbers $(x, y) \in [0,1]^2$

$$|x - y| < \varepsilon \implies \left|x^2 - y^2\right| < 2\varepsilon. \tag{4.20}$$

We define $p_{0/1}$ the probabilities of measuring $U(x)\,|0\rangle$ in state $|0\rangle$ and $|1\rangle$ respectively. Recalling that

$$|\langle 1|\, U(x)\,|0\rangle - f(x)| < \varepsilon/4, \tag{4.21}$$

we can write

$$\left|p_1 - f(x)^2\right| < \varepsilon/2 \tag{4.22}$$

$$\left|p_0 - (1 - f(x)^2)\right| =$$

$$\left|1 - |\langle 1|\, U(x)\,|0\rangle|^2 - (1 - f(x)^2)\right| < \varepsilon/2. \tag{4.23}$$

Finally,

$$|\langle Z\rangle - g(x)| \leq \left|p_1 - f(x)^2\right| + \left|p_0 - (1 - f(x)^2)\right| < \varepsilon. \tag{4.24}$$

This yields the uniform density of quantum functions in the set of bounded continuous functions.

From uniform density on continuous functions to pointwise density on discontinuous functions

We first start by highlighting the difference between uniform convergence and pointwise convergence and illustrate it with an example.

Definition 4.6 (Pointwise convergence)
Let $\{f_k|f_k : \mathcal{X} \to \mathbb{R}\}_{1\leq k\leq\infty}$ be a sequence of functions, and let $f : \mathcal{X} \to \mathbb{R}$ be another function defined on the same domain $\mathcal{X}$. We say that the sequence $\{f_k\}$ converges pointwise to f if, for each $x \in \mathcal{X}$, the sequence of real numbers $\{f_k(x)\}_{1\leq k\leq\infty}$ converges to $f(x)$ as k approaches infinity,

$$\lim_{n\to\infty} f_k(x) = f(x) \quad \textit{for all } x \in \mathcal{X}.$$

Definition 4.7 (Uniform convergence)
Let $\{f_k|f_k : \mathcal{X} \to \mathbb{R}\}_{1\leq k\leq\infty}$ be a sequence of functions, and let $f : \mathcal{X} \to \mathbb{R}$ be another function defined on the same domain $\mathcal{X}$. We say that the sequence f_k converges uniformly to f if, for any given $\varepsilon > 0$, there exists an $K \in \mathbb{N}$ such that for all $k \geq K$ and for all $x \in \mathcal{X}$, the difference $|f_k(x) - f(x)|$ is less than ε,

$$\forall\varepsilon > 0, \exists K \in \mathbb{N} : \forall k \geq K, \forall x \in \mathcal{X}, |f_k(x) - f(x)| < \varepsilon.$$

Uniform convergence is stronger, it implies pointwise convergence, but the reverse is not true. For example, consider the step function f.

$$f(x) = \begin{cases} -1, & \text{if } x \in [-1, 0[\\ +1, & \text{if } x \in [0, +1] \end{cases} \tag{4.25}$$

It is impossible to define a sequence of continuous functions that would uniformly converge to it, however, it is possible to have a sequence of continuous functions that converges pointwise to it, see Figure 4.6.

$$f_k(x) = \begin{cases} -1, & \text{if } x \in [-1, 1/k[\\ 1 + kx, & \text{if } x \in [-1/k, 0] \\ +1, & \text{if } x \in]0, +1] \end{cases} \tag{4.26}$$

In [145] Baire defined hierarchical pointwise convergence classes of functions. Baire class 0 is the set of continuous functions, and the class c is the set of functions that are the pointwise limit of class $c-1$. In particular in [146, 147], it was proven that bounded piece-wise continuous functions are of class 1. This means that for any bounded piece-wise continuous function f there exists a sequence of bounded continuous functions that converges pointwise to f. This means that bounded continuous functions are dense in bounded piece-wise continuous functions in the pointwise topology. Therefore, building on Theorem 4.6 we have the following theorem, writing $\mathcal{B}$ as the set of piecewise continuous functions.

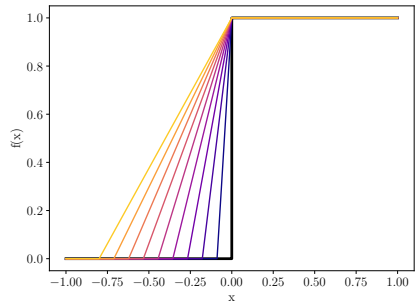

Figure 4.6.: A sequence of continuous functions converging pointwise but not uniformly to the step function.

Theorem 4.7
For any function $f : \mathcal{B}([0,1]^M \to [-1,1])$*, there exists a sequence (indexed by* k*) of one qubit circuit and observables with unit spectral norm,*

$$\{(U_k(x) : [0,1]^M \to \mathcal{U}(2), O_k)\}_{1 \leq k \leq \infty} \tag{4.27}$$

such that the sequence of functions $\{g_k\}$ *with*

$$g_k(x) = \langle 0 | U_k^\dagger(x) O_k U_k(x) | 0 \rangle \tag{4.28}$$

converges pointwise to f*.*

Construction of the observable dense encoding circuit

Note: We have used slight abuse of notations in the below proof to increase readability, specifically in the approximations noted as $\equiv_\varepsilon$.

Considering architecture such as in [148] and universal one qubit gate as $R_x(\alpha)R_z(\beta)R_x(\gamma)$, for any number of qubits, it is possible to design a circuit $U : [0, 2\pi)^L \to \mathcal{U}(2^n)$ made only of a finite number of fixed gates (CNOT and constant rotations) and parameterized σ_z rotations gates that can reach any pure quantum state when applied to state $|0\rangle$,

$$\forall |\psi\rangle, \exists \theta \in [0, 2\pi)^L, |\psi\rangle = U(\theta) |0\rangle . \tag{4.29}$$

Let's consider a distribution p_ψ over pure states. Because the architecture above can reach any pure state, there exists a corresponding distribution over the parameters p_θ such that the distribution $V(\theta) |0\rangle, \theta \sim p_\theta$ matches

perfectly p_ψ in distribution. We define $g : [0,1]^L \to [0,2\pi)^L$ as the mapping that transforms the uniform distribution over $[0,1]^L$ into p_θ. We have $V(f(X))\left|0\right\rangle, X \sim U([0,1]^L)$ matches perfectly p_ψ in distribution, where V is composed of a finite number of fixed gates and L σ_z rotation gates parameterized by $g_l(X)$.

Lemma 4.5
For any distribution over pure states p_ψ, there exists a circuit architecture V made of constant gates and parameterized σ_z rotations, and a mapping g such that

$$X \sim U([0,1]^L), V(g(X))\left|0\right\rangle \sim p_\psi \tag{4.30}$$

Next, we decompose $R_z \circ g$ gates into a sequence of constant gates and parameterized σ_z rotations.

From [25], in the proof in the appendix, it is shown that there exists a quantum circuit W taking a multidimensional input and approximating the following parameterized quantum gate,

$$\forall f : [0,1]^L \to [0,1], \phi : [0,1]^L \to [0,2\pi), \exists W,$$
$$W(x) \equiv_\varepsilon \begin{bmatrix} \sqrt{1-f(x)^2}e^{+i\phi(x)} & -f(x)e^{+i\phi(x)} \\ f(x)e^{-i\phi(x)} & \sqrt{1-f(x)^2}e^{-i\phi(x)} \end{bmatrix}. \tag{4.31}$$

Choosing $\phi = 0$, and $f = \sin g$, we have $\sqrt{1-f^2} = \cos g$, and we note that $W(x) = R_z \circ g$. We can conclude the following.

Lemma 4.6

$$\forall f : [0,1]^L \to [0,1], \exists W, R_z \circ f \equiv_\varepsilon W, \tag{4.32}$$

where W is a quantum circuit made of constant gates and R_z gates applied to individual components of x.

Combining both above lemmas we get the following.

Lemma 4.7
For any distribution over pure states p_ψ, there exists a circuit architecture V made of constant gates and parameterized z rotations such that

$$X \sim U([0,1]^L), V(X)\left|0\right\rangle \sim_\varepsilon p_\psi. \tag{4.33}$$

We are now going to use that lemma to prove that n-qubits expectation value samplers are universal for $\exp(n)$-dimensional distributions with constant support if the observables are allowed to have $\exp(n)$ norms.

We are given an arbitrary random variable Y following a M-dimensional distribution p_Y with support $[-1,1]^M$. We define the following state over n qubits with $M = 2^n - 1$.

$$|\psi(Y)\rangle = \sum_{m \leq M} Z_m |m\rangle + \sqrt{1 - \sum_{m \leq M} Z_m^2} |M+1\rangle \tag{4.34}$$

$$Z_m = \sqrt{\frac{Y_m + 1}{2M}} \tag{4.35}$$

We define as p_ψ as the probability density functions over states when $Y \sim p_Y$, using the previous lemma we get a circuit W composed only of constant gates and z rotations gates with one of the L parameters as input that approximates p_ψ. We define the observables $\forall m \leq M, O_m = 2M |m\rangle \langle m| - I$. They have spectral norm $\|O_m\| = 2M - 1 = \Theta(2^n)$. In addition, $\langle \psi(Y)| O_m |\psi(Y)\rangle = Y_m$. Therefore, the expectation value sampler $(W, O, U([0,1]^L))$ approximates p_Y. This concludes the proof to Theorem 2.

4.5.3. Proof of necessary resources for universality

In this appendix, we prove Theorem 3 about the necessary resources for EVS to achieve universality, which we recall below.

Theorem
For an n-qubit expectation value sampling model $(U_\theta, \mathbf{O}, p_X)$ to be able to approximate any distribution with support in $[-1,1]^M$ to any accuracy $\varepsilon > 0$ with respect to the Wasserstein distance, it is necessary that for every $m \leq M$:

1. $\lambda_{\min}(O_m) \leq -1 + \varepsilon$ *and* $\lambda_{\max}(O_m) \geq +1 - \varepsilon$
2. $n \in \Omega\left(\frac{M(1-\varepsilon)^2}{\Lambda(O_m)}\right) \subseteq \Omega\left(\frac{M(1-\varepsilon)^2}{\|O_m\|^2}\right)$

with $\lambda_{\min/\max}(O)$ returning respectively the minimum and maximum eigenvalues of observable O, and $\Lambda(O) \coloneqq -\lambda_{\min}(O)\lambda_{\max}(O)$.

Proof: Let's suppose there is an n-qubit expectation value scheme with M observables $\mathbf{O}$ that is able to approximate any distributions with support included in $[-1,1]^M$ to ε with respect to the first Wasserstein distance W_1.

Because the expectation value model is universal, it means that for any vertex of the hypercube $c \in \{-1,1\}^M$, it can approximate the Dirac delta at c. We then use the lemma below.

Lemma 4.8
Any distribution that is ε-close (in the Wasserstein distance) to the Dirac delta at a given point must have nonzero support within the Euclidean distance sphere centred in that point with radius ε.

This means that there is non-zero support on the ε sphere around c, and therefore there exists a quantum state ρ_c whose list of expectations is ε-close to that point. This yields the following result.

$$\forall c \in \{-1,1\}^M, \exists \rho_c, \sum_{m=1}^{M} (\mathrm{Tr}(O_m \rho_c) - c_m)^2 \leq \varepsilon^2. \tag{4.36}$$

Because the above is a sum of positive components, we can write $\forall m$

(a) if $c_m = 0$, then $-1 - \varepsilon \leq \mathrm{Tr}(O_m \rho_c) \leq -1 + \varepsilon$

(b) if $c_m = 1$, then $+1 - \varepsilon \leq \mathrm{Tr}(O_m \rho_c) \leq +1 + \varepsilon$

The above yields conditions on the spectrum of O_m. We note $\lambda_{\min}$ and $\lambda_{\max}$ respectively the minimum and maximum eigenvalues of O_m. We define $\gamma \coloneqq 1 - \varepsilon$. We know that $\forall \rho, \mathrm{Tr}(O\rho) \geq \lambda_{\min}$ therefore, $-\gamma \geq \lambda_{\min}$, the same reasoning applies for the maximum eigenvalue, yielding:

(a) $\lambda_{\min} \leq -\gamma$,

(b) $\lambda_{\max} \geq +\gamma$.

For the rest of the proof, we combine approaches from the proof of theorem 2.6 in [149] and that of B.1 in [150]. We define the two-outcome POVMs $\{E_m, I - E_m\}$ with

$$E_m = \frac{O_m - \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}} \tag{4.37}$$

We define $\beta \coloneqq \frac{-\lambda_{\min}}{\lambda_{\max} - \lambda_{\min}}$. The spectral inequalities yield $\beta \geq \frac{1}{2}$. With this definition, the two conditions above translate to:

(a) $c_m = 0, p_0 \coloneqq \mathrm{Tr}(E_m \rho_c) \leq \beta - \frac{\gamma}{\lambda_{\max} - \lambda_{\min}}$

(b) $c_m = 1, p_1 \coloneqq \mathrm{Tr}(E_m \rho_c) \geq \beta + \frac{\gamma}{\lambda_{\max} - \lambda_{\min}}$

We define the amplified Positive Operator-Valued Measures (POVMs) that apply $\{E_m, I - E_m\}$ to $L \geq 1$ copies of ρ and return 1 if and only if at least βL copies of the original POVMs return 1.

From Holevo's bound (found as Theorem 5.1 in [138]), for the amplified scheme to correctly identify the corresponding bit $b_m = (c_m + 1)/2$ with probability q it is necessary that

$$nL \geq (1 - H(q))M, \tag{4.38}$$

where H is the binary entropy function.

We define the random variable $X_{m,l}$ which takes the value of the output of the POVM of the m-th observable on the l-th copy. We define $X_m^{(L)} := \frac{1}{L}\sum_l X_{m,l}$.

In the case $c_m = -1$, we have $\mathbb{E}[X_m^{(L)}] = p_0$. The probability of the amplified POVMs yielding the wrong output is

$$P(X_m^{(L)} > \beta) \leq P\left(X_m^{(L)} > p_0 + \frac{\gamma}{\lambda_{\max} - \lambda_{\min}}\right) \tag{4.39}$$

Recalling that $\beta \geq 1/2$, we can use the Chernoff bound on the Bernoulli variable $X_m^{(L)}$ and we get

$$P(X_m^{(L)} > p_0 + \frac{\gamma}{\lambda_{\max} - \lambda_{\min}}) \leq \exp -\frac{\gamma^2 L}{2(-\lambda_{\min})\lambda_{\max}} \tag{4.40}$$

We define $\Lambda = (-\lambda_{\min})\lambda_{\max}$, we have $\gamma^2 \leq \Lambda \leq \|O\|^2$.

For the probability of the amplified POVMs to yield the correct output with probability q it is necessary that $P(X_m^{(L)} \leq \beta) \geq q$.

Finally, we get

$$\log(1/q) \leq \frac{\gamma^2 L}{2\Lambda} \tag{4.41}$$

Combining Chernoff's and Holevo's inequalities, we conclude the proof of Theorem 3:

$$\Lambda \geq \frac{1 - H(q)}{\log(1/q)} \frac{\gamma^2 M}{n}. \tag{4.42}$$

Note: The above is a tighter condition than in Theorem 2.6 in [150] but falls back to it, when $\Lambda = \|O\|^2$, which corresponds to $\lambda_{\min} = -\|O\|$ and $\lambda_{\max} = \|O\|$. In the opposite scenario, we have $\Lambda = \gamma^2$, which corresponds to constant norm observables, yielding $n \in \Omega(M)$. The norm of observables affects the number of measurements to reach a desired additive accuracy.

Note: The above necessary conditions use results which involve a more general case where we assume that prior to measurement we use a general parameterized quantum channel which can also prepare mixed states.

However, we can reduce this to the special case of unitaries as well. By purification, mixed states can be mimicked by using $2n$ qubits pure states, since we are only interested in scalings, the factor 2 plays no role in the second condition of Theorem 3, and the necessary conditions also hold for pure states.

4.5.4. Approximation of expectation values

In this appendix, we justify the connection between the spectral norm of observables and the number of measurements. In practice, we do not have access to exact expectation values and we have to estimate them through sampling. This creates a distribution $p_{\hat{Y}}$ (n dimensional) slightly different from the distribution with exact expectation values p_Y. For simplicity, we assume that shot noise $\rho_\varepsilon \sim \mathcal{N}(0, \varepsilon I_n)$ (n dimensional) acts like an additional Gaussian noise with zero mean and standard deviation ε. ρ_ε and Y are independent and we consider the random variable $\hat{Y} = Y + \rho_\varepsilon$. Therefore the density $p_{\hat{Y}}$ is the convolution of p_Y and p_{ρ_ε}. Using Lemma 7.1.10 from [151], we know that the Wasserstein distance between p_Y and $p_{\hat{Y}}$, $W_p(p_Y, p_{\hat{Y}}) \in O(\varepsilon)$.

Lemma 4.9
An arbitrary expectation value sampler outputs an M-dimensional random vector $Y \sim p_Y$ with an infinite number of measurements, i.e. with access to exact expectation values. We consider the same circuit but with a finite number of measurements T that estimates expectation value by sampling and averaging for each observable and yields a random vector $\hat{Y} \sim p_{\hat{Y}}$. The number of measurements T required to guarantee that the Wasserstein distance between both distributions is smaller than ε satisfies

$$T \in \Theta\left(\frac{M\|O\|}{\varepsilon^2}\right). \tag{4.43}$$

In practice, different techniques exist to estimate expectation values of observables with different degrees of measurement efficiency, with shadow tomography techniques surpassing the "vanilla estimation". For simplicity, we consider a vanilla estimation where each observable with norm $\|O\|$ is measured t times and the average is returned. This yields a shot noise close to the Gaussian model above with $\varepsilon^2 \in \Theta(\|O\|/t)$. The total number of measurements T is then $T = tM$, which yield $T \in \Theta(M\|O\|/\varepsilon^2)$

Note that the first result cannot be trivially applied to techniques such as shadow tomography. Indeed the corresponding shot noise ρ cannot

in general be modeled by a Gaussian independent noise, at the least the covariance matrix will in general not be proportional to the identity.

4.5.5. Additional expressivity tools

In this appendix, we provide additional tools to analyze the expressivity of expectation value samplers. Specifically the choice of observabes, and the choice of random variable encoding.

Primary mapping and the choice of observables

We are using the standard Pauli basis for the space of $2^n \times 2^n$ Hermitian operators $\mathbf{P}$. It is composed of all possible combinations of n Pauli matrices $\sigma_{0,1,2,3}$, which yields $|\mathbf{P}| = 4^n$. We formalize as follows

$$\begin{aligned}\mathbf{P} :=& (P_k, k \in \{0,1,2,3\}^n) && (4.44)\\ =& (\otimes_{1\leq i\leq n}\sigma_{k_i}, k_i \in \{0,1,2,3\}). && (4.45)\end{aligned}$$

Any vector of M observables $\mathbf{O} = (O_m)$, can be expressed as a linear mapping applied on the vector of all Pauli strings: $\mathbf{O} = A\mathbf{P}$, where A is an $M \times 4^n$ matrix, and $\mathbf{P}$ is a 4^n dimensional vector. Therefore the distribution associated with the Pauli basis encompasses any distributions, which leads us to define the primary mapping as follows.

Definition 4.8 (Primary mapping)
The primary mapping g of an n-qubit encoding circuit $U_\theta(x)$ is defined as the mapping of the associated expectation value sampling model with the Pauli basis $\mathbf{P}$, defined as $(U_\theta(x), \mathbf{P}, p_X)$ according to the definition in the main body. It can be expressed as follows

$$x \in [0, 2\pi)^N \xrightarrow{g} \left(\langle 0| U_\theta(x)^\dagger P_k U_\theta(x) |0\rangle\right)_{1\leq k\leq 2^n}. \tag{4.46}$$

It yields the 4^n-dimensional random variable $Z = g(X)$ when $X \sim p_X$.

It is easy to see that any distribution obtained by an expectation value sampling model can always be expressed by considering an intermediary output of the 4^n set of observables, followed by the linear mapping A. We capture this idea in the following theorem.

Lemma 4.10
Given an encoding circuit $U_\theta(x)$ and random variable with distribution p_X, for any choice of observables $\mathbf{O}$, the expectation value sampling model

$(U_\theta(x), \mathbf{O}, p_X)$ is a linear transformation of the expectation value sampling model $(U_\theta(x), \mathbf{P}, p_X)$, where $\mathbf{P}$ is the Pauli basis per definition of the primary mapping in Definition 4.8.

This concept of primary mapping has immediate consequences on the possible correlation of output variables of expectation value sampling models. For a given data encoding part on n qubits, the primary mapping will yield a random variable Z with a covariance matrix C_z. Because any expectation value sampling model based on the same data encoding is a linear transformation of the primary mapping, the number of uncorrelated variables is limited by the number of non-null eigenvalues of the covariance matrix C_z, which is upper bounded in any case by $4^n - 1$.

Lemma 4.11
Given an encoding circuit $U(x)$ over n qubits and random variable with distribution p_X, we note L the number of non-zero eigenvalues of the covariance matrix of the primary mapping. For any choice of observables, any expectation value model using $(U(x), p_X)$ will yield at most L uncorrelated variables. In addition $L \leq 4^n - 1$.

Random variable encoding as a polynomial chaos expansion

After focusing on the choice of observables, in this subsection, we analyze the impact of the choice of the input random variable and the circuit encoding on the expressivity. We propose the polynomial chaos expansion [152] as a useful tool to analyze the expressivity of expectation value sampling models, as the analogue of the Fourier decomposition. The general polynomial chaos expansion is a representation of random variables as a vector in a Hilbert space of orthogonal functions, as defined below.

Definition 4.9 (Generalized Polynomial Chaos Expansion)
A generalized chaos expansion is characterized by a probability density function p_X defined on the support $\mathcal{X} \subseteq \mathbb{R}^M$ with finite moments (usually chosen as standard distributions, such as Gaussian or uniform). This choice defines an inner product for functions in $\{f : \mathcal{X} \to \mathbb{R}\}$:

$$\langle f|g\rangle_{p_X} \coloneqq \int_{\mathcal{X}} f^*(x)g(x)p_X(x)dx. \tag{4.47}$$

This choice of inner product comes with the choice of an ordered family of functions, usually polynomials, that are orthonormal with respect to the above inner product.

$$\Phi_{p_X} = \{\phi_l : \mathcal{X} \to \mathbb{R}, \forall(k,l), |\phi_l\rangle \langle\phi_k| = \delta_{k,l}\} \tag{4.48}$$

A generalized chaos expansion is a representation of a random variable Y with probability density p_Y as a vector α in this Hilbert space, such that:

$$Y = \sum_{l=0}^{\infty} \alpha_l \phi_l(X) \sim p_Y, X \sim p_X \tag{4.49}$$

This provides a Hilbert space as a potential structure to study random variable mappings. In particular, one of the main results of polynomial chaos expansion is that they are universal generative models.

Common pairs of distribution and associated orthogonal polynomials family can be found in Table 4.1. In the context of expectation value sampling, we focus on the family of functions orthonormal with respect to the inner product associated with the uniform distribution on $\mathcal{X} = [0, 2\pi)^M$

$$\Phi = \{ \prod_{1 \leq m \leq M} e^{ik_m x_m}, k \in \mathbb{Z}^M \} \tag{4.50}$$

Quantum reuploading circuits are a widely used class of parameterized quantum circuits that output a function of the data. They are used in a regressive context, where optimization techniques are used such that their output fits a target function. It has been widely studied and used that if they use integer-valued spectrum Hamiltonian, their output hypothesis function can be decomposed as an exact finite Fourier series [26]. This means that there exists $c_{\mathbf{k},l} \in \mathbb{C}$ such that the hypothesis function f can be exactly written as a finite Fourier series:

$$f(x) = \sum_{k_0, k_1, \cdots, k_M = -K}^{+K} c_{\mathbf{k},l} \prod_{1 \leq m \leq M} e^{ik_m x_m}, . \tag{4.51}$$

This fact extends to a generative modelling context where expectation value sampling models using integer-valued quantum reuploading circuits yield distributions with an exact finite polynomial chaos expansion. We formalize this below.

Theorem 4.8
Any expectation value sampling model using a quantum reuploading model with integer-valued spectrum $U_\theta(x)$, together with the uniform distribution on $[0, 2\pi)^M$ outputs a random variable Y that has an exact finite polynomial chaos expansion for any choice of observables.

This subsection formalizes the tight connection between quantum circuits used in a regressive context with their use in a generative context. Therefore

Normal distribution	Hermite polynomials
Uniform distribution	Legendre polynomials
Exponential distribution	Laguerre polynomials
Beta distribution	Jacobi polynomials

Table 4.1.: Pairs of distributions and corresponding orthonormal families commonly used in General Polynomial Chaos expansion.

it is expected that, beyond the universality, many properties of such models can be transferred from a regressive context to a generative context, but we leave that for future work.

Choice of input distribution

Lastly, we discuss the choice of input random variable, noted as p_X. We highlight that the universality theorems in this chapter apply to the uniform distribution. This, together with generalized chaos expansion considerations makes the uniform distribution a natural choice for the input random variable of expectation value samplers. This is due mostly to the fact that it has a bounded support. Indeed, it is known that the most commonly used parameterized quantum circuits are periodic, in particular when they have a finite Fourier decomposition. This is also why the universality of quantum reuploading circuits is proven for functions on bounded domains, corresponding to a half period. In contrast, let us consider an expectation value sampler $(U, \mathbf{O}, p_X)$, where the input random variable follows a Gaussian distribution $X \sim \mathcal{N}(0, 1)$, which has unbounded support and for which the mapping f is 1-periodic. $X = 0$ is the highest probability event and will yield the same output as a very low probability event, for example, $X = 100$. This means that a very low probability input event and a very high probability input event will yield the same output sample, which is a feature rarely considered desirable. This choice of uniform distribution is in contrast to classical GANs where Gaussian distributions are typically preferred.